AncientBench: Towards Comprehensive Evaluation on Excavated and Transmitted Chinese Corpora

Zhihan Zhou[1,2], **Daqian Shi**[2,3], **Rui Song**[1,2], **Lida Shi**[2,4], **Xiaolei Diao**[*1,2,5], **Hao Xu**[*1,2]

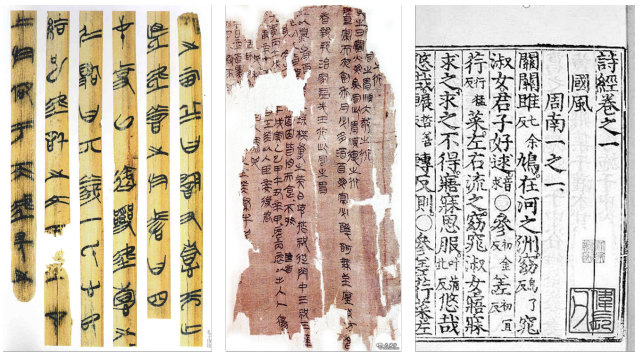
[1]College of Computer Science and Technology, Jilin University
[2]Key Laboratory of Ancient Chinese Script, Culture Relics and Artificial Intelligence, Jilin University
[3]Digital Environment Research Institute, Queen Mary University of London
[4]School of Artificial Intelligence, Jilin University
[5]Department of Information Engineering and Computer Science, University of Trento
{zhzhou25, shild21}@mails.jlu.edu.cn, d.shi@qmul.ac.uk, {songrui, xuhao}@jlu.edu.cn, xiaolei.diao@unitn.it

Abstract

Comprehension of ancient texts plays an important role in archaeology and understanding of Chinese history and civilization. The rapid development of large language models needs benchmarks that can evaluate their comprehension of ancient characters. Existing Chinese benchmarks are mostly targeted at modern Chinese and transmitted documents in ancient Chinese, but the part of excavated documents in ancient Chinese is not covered. To meet this need, we propose the AncientBench, which aims to evaluate the comprehension of ancient characters, especially in the scenario of excavated documents. The AncientBench is divided into four dimensions, which correspond to the four competencies of ancient character comprehension: glyph comprehension, pronunciation comprehension, meaning comprehension, and contextual comprehension. The benchmark also contains ten tasks, including radical, phonetic radical, homophone, cloze, translation, and more, providing a comprehensive framework for evaluation. We convened archaeological researchers to conduct experimental evaluations, proposed an ancient model as baseline, and conducted extensive experiments on the currently best-performing large language models. The experimental results reveal the great potential of large language models in ancient textual scenarios as well as the gap with humans. Our research aims to promote the development and application of large language models in the field of archaeology and ancient Chinese language.

Introduction

As one of the most spoken languages in the world, Chinese has received extensive attention in recent years in the field of natural language processing (NLP). As an important part of the Chinese language, ancient Chinese carries extremely rich historical and cultural information, and its study is of great significance for the traceability of Chinese history, the protection of cultural heritage, and the development of historical linguistics. However, traditional ancient Chinese research methods are highly dependent on the researcher's memory and linguistic intuition, which is often inefficient and difficult to deal with large-scale corpus systematically, limiting the breadth and depth of research.



(a) (b) (c)

Figure 1: Comparison of excavated documents with transmitted documents. (a) Bamboo Book of Chu, excavated documents. (b) Silk Books, excavated documents. (c) Book of Poetry, transmitted documents.

With the rapid development of artificial intelligence technology, especially large language models (LLMs), ancient Chinese research is gradually stepping towards a new paradigm of data-driven and model-supported. LLMs have demonstrated excellent generalization ability in natural language understanding and natural language generation, and have shown great potential in text analysis, language structure modeling, and so on. In order to systematically evaluate the capability of these models, academics have gradually constructed a series of standardized evaluation systems, such as MMLU(Hendrycks et al. 2021a), BIG-bench(Srivastava et al. 2022), and HELM(Bommasani, Liang, and Lee 2023), etc., which have become important tools for measuring the general language intelligence of models. This technological trend offers new opportunities for ancient Chinese processing, especially in understanding difficult ancient texts.

In the Chinese context, based on its linguistic characteristics and application requirements, a series of benchmarks have emerged to evaluate the capability of Chinese LLMs, such as CLUE(Xu et al. 2020a), CMMLU(Li et al. 2024),

*Corresponding authors

and MMCU(Zeng 2023), etc. Most of these benchmarks cover a wide range of domains in Chinese scenarios, such as language comprehension, logical reasoning, instruction execution, etc., which promote the evaluation and iterative optimization of Chinese language models. In recent years, some datasets related to ancient Chinese have also appeared, e.g., ACLUE(Zhang and Li 2023) and WYWEB(Zhou et al. 2023). However, the tasks related to ancient Chinese in such benchmarks are scarce, mostly limited to syntax parsing, with no unified evaluation criteria, making it hard to fully assess the comprehension ability of language models for ancient Chinese.

Ancient Chinese documents are mainly divided into two categories: transmitted documents (TraDoc) and excavated documents (ExcDoc) (Guiyuan 2023). Transmitted documents refer to those that have been handed down from ancient times through hand-copying, engraving, copying, etc., such as the Analects, the Grand Scribe's Records, and the Book of Poetry. These documents have been continuously collated and annotated by future generations, forming a relatively stable textual system with standardized language and clear structure, which facilitates the construction of standardized comprehension and reasoning tasks, and is the main source of corpus for most of the current ancient literature evaluation tasks. In contrast, excavated documents refer to ancient written materials obtained through archaeological excavations, mainly including oracle bones, Bronze Inscriptions, Bamboo Book of Chu, Silk Books, and so on. These documents are much earlier in time and have a more primitive and authentic language style, directly reflecting language use and culture in ancient societies. We give a detailed comparison between transmitted documents and excavated documents in Figure 1.

On the one hand, excavated documents are mostly obtained from original and unearthed objects (Chi et al. 2022) that have not been copied, compiled, or politically interfered with over the generations, thus preserving the most primitive form and content of the text (Boltz 1986). On the other hand, excavated documents usually have clear archaeological layers, coexisting artifacts, and scientific dating data, which can accurately locate their age (Jane 2014). Therefore, excavated documents have stronger originality, authenticity, and clarity of time and region than transmitted documents. Excavated documents are of irreplaceable importance in the study of ancient Chinese for the understanding of ancient documents. Existing LLMs in the field of ancient Chinese still mainly rely on transmitted documents in the training and evaluation process, and the coverage of excavated documents is extremely limited because most of the texts of excavated documents are concentrated on the surface of tortoise shells and bamboo slips, which are difficult to access and covered with complex noises (Shi et al. 2022b). Some attempts include the use of OCR technology to detect and recognize excavated documents (Diao et al. 2023b; Yue et al. 2025), but only a small amount of data is available because the existing text encoding in computers cannot cover all excavated documents. The above difficulties have hindered the research of LLMs in the field of ancient Chinese.

To bridge this gap, we propose AncientBench, a benchmark for evaluating the comprehension of ancient texts, especially excavated documents, in large language models. From the perspective of data corpus, AncientBench mainly adopts the pre-Qin period, i.e., the Xia dynasty, Shang dynasty, Zhou dynasty, Spring and Autumn periods, and Warring States period, focusing mainly on Oracle Bone Inscriptions, Bronze Inscriptions, and Bamboo Book of Chu. From the perspective of the competencies evaluated and the design of the questions, the AncientBench examines the four competencies of LLMs in the field of ancient Chinese, i.e., glyph comprehension, pronunciation comprehension, meaning comprehension, and contextual comprehension. The AncientBench contains ten tasks, i.e., Radical, Radical Meaning, Pronunciation, Phonetic Radical, Homophone, ExcDoc Word, TraDoc Word, Cloze, Phonetic Loan Character, and Translation. We aim to comprehensively and objectively evaluate the ability of LLMs to understand ancient texts.

We convened researchers from the interdisciplinary fields of archaeology and artificial intelligence to conduct experiments based on the AncientBench to evaluate the comprehension of ancient characters as humans, and evaluated LLMs with zero-shot and few-shot, respectively. The experimental results show that there is still a gap between their comprehension of ancient texts and that of humans, and most of the models still need to improve their comprehension of ancient texts. Our study aims to propose a comprehensively and multi-dimensional benchmark in the field of ancient Chinese, with the aim of promoting the research of LLMs in the field of ancient Chinese. Our contribution is summarized below.

- We have achieved the digitization of ancient characters. We propose a three-stage method for digitization. With this method, we can process excavated documents, which facilitates the training and evaluation of ancient LLMs.
- We construct AncientBench. For the first time, we introduce excavated documents into the field of natural language processing, construct AncientBench focusing on excavated documents, and propose a novel and comprehensive evaluation benchmark for the research of LLMs in the field of ancient Chinese.
- We convened archaeological researchers to evaluate the comprehension of ancient characters in humans. At the same time, we proposed an ancient model and conducted extensive experiments based on AncientBench.

Related Work

The development of benchmarks for evaluating models has been an important research topic in NLP. Early models mainly relied on pretraining fine-tuning methods, and a series of benchmarks have been proposed for NLU tasks, such as SentEval(Conneau and Kiela 2018), GLUE(Wang et al. 2018) and SuperGLUE(Wang et al. 2019), where CLUE(Xu et al. 2020b) contains more than 10 tasks covering most NLP problems. With the development of LLMs in recent years, A series of benchmarks have been proposed for NLG tasks, such as Sakaguchi(Sakaguchi et al. 2021), Hendrycks(Hendrycks et al. 2021b), and Austin(Austin et al. 2021). MMLU(Hendrycks et al. 2021a) is extremely popular

Competencies	Task	Questions	Description
Glyph	Radical	8438	Radical Recognition
	Radical Meaning	1432	Radical Meaning Recognition
Pronunciation	Pronunciation	3886	Pronunciation Recognition
	Phonetic Radical	304	Phonetic Radical Recognition
	Homophone	2265	Homophone Recognition
Meaning	ExcDoc Word	365	Excavated Documents Word Understanding
	TraDoc Word	1504	Transmitted Documents Word Understanding
Contextual	Cloze	4875	Cloze Task for Excavated Documents
	Phonetic Loan Character	4637	Phonetic Loan Character Recognition
	Translation	1001	Translation

Table 1: Framework of AncientBench. **Competencies** denotes the four competencies examined by the benchmark, e.g. Glyph denotes glyph comprehension. **Task** denotes the ten tasks. **Questions** denotes the number of questions included in each task. **Description** denotes the description of the task.

in LLM evaluations due to its standardized question format, comprehensive coverage, and uniformity of the evaluation criteria. However, most of the benchmarks mentioned above focus on evaluating in English.

As the most spoken language in the world, Chinese has a very important research significance. Xu proposed CLUE(Xu et al. 2020b) and SuperCLUE(Xu et al. 2023), which cover most of the tasks of Chinese NLU. After MMLU(Hendrycks et al. 2021a) was proposed, many researchers proposed Chinese scenarios based on the MMLU evaluation framework. Multi-domain and multidimensional benchmarks, such as MMCU(Zeng 2023), CMMLU(Li et al. 2024), and AGIEval(Zhong et al. 2024), where MMCU covers four major domains and CMMLU covers more than 20 multilingual and Chinese languages. In addition, CG-Eval(Zeng et al. 2024b) and M3KE(Liu et al. 2023) used multitask multiple question types to evaluate LLMs.

In recent years, benchmarks for evaluating ancient Chinese have made rich progress in several directions. (Zinin and Xu 2020) further open-sourced 20 historical traveling materials and other ancient corpus, enriching the diversity of the domain; FSPC(Shao et al. 2021) and CcMP(Li et al. 2021) focus on the comprehension task of ancient poems, and the CUGE(Yao et al. 2022) focuses on the poetry matching subtask based on CcMP. In addition, (Pan et al. 2022), (Wang and Ren 2022), and (Liu et al. 2022) have successively introduced task sets covering syntactic analysis, topic mining, and sentiment classification; Comprehensive benchmarks such as CCLUE(Yao et al. 2022) and WYWEB(Zhou et al. 2023) integrate various tasks from text classification to poetry analysis and machine reading comprehension. ACEVAL(Wei et al. 2024) is mainly targeted at LLMs, and integrates a variety of tasks and datasets, constructing a comprehensive and integrated benchmark for ancient Chinese.

AncientBench

AncientBench Overview

AncientBench consists of four dimensions of competency evaluations and ten tasks, as shown in Table 1. The four dimensions of competency evaluation include glyph comprehension, pronunciation comprehension, meaning comprehension, and contextual comprehension. The ten tasks include Radical, Radical Meaning, Pronunciation, Phonetic Radical, Homophone, ExcDoc Word, TraDoc Word, Cloze, Phonetic Loan Character, and Translation. In order to ensure the quality of the competency dimensions and task definitions, we used the following criteria for the definitions.

The first is authority and credibility. We invited experts in the field of archaeology to analyze the comprehension of ancient characters from a human perspective, i.e., what are the foundational abilities that persons need if we consider them to have the ability to understand ancient texts. Combining archaeologists' analysis and classical tasks in the field of natural language processing, we finally identified these four dimensions.

The second is diversity and comprehensiveness. In order to ensure the comprehensiveness of the AncientBench, we defined it from two perspectives: the characteristics of ancient Chinese documents and the difficulty of textual comprehension. The four dimensions of the AncientBench represent a surface-to-surface understanding of an ancient Chinese character, where glyph understanding and pronunciation understanding represent the external features of a character. The glyph is key to transferring the semantics of characters (Shi et al. 2022a). In glyph comprehension, Radical Recognition refers to the recognition of the components of a character (Diao et al. 2023a), and Radical Meaning refers to the recognition towards the meaning of the components. In pronunciation comprehension, Pronunciation Recognition refers to the recognition of the pronunciation of a character, and we refer to the notation in the ancient Chinese dictionary ShuoWen(Xu 1981) as the definition of character pronunciation, while Phonetic Radical Recognition refers to the recognition of the component of a character that is responsible for the pronunciation of the character (Ho, Ng, and Ng 2003). Meaning comprehension and contextual comprehension are the internal meanings of a character (Nelson and Stage 2007). In meaning comprehension, ExcDoc Word refers to the understanding of the word meanings in excavated documents, and TraDoc Word refers to the understanding of the word meanings in transmitted documents. Con-

textual comprehension involves identifying Phonetic Loan Characters (Yang 2000) in excavated texts. AncientBench's four tasks vary in difficulty: glyph and pronunciation comprehension is at the character level, meaning comprehension is at the word level, and contextual comprehension is at the sentence level.

Finally, the standardization of the criteria. In order to ensure the uniformity and certainty of the evaluation criteria of AncientBench. We chose the classic tasks in the field of natural language processing, such as cloze, translation, etc. as the evaluation tasks. Then we further processed these tasks and finally displayed them in multiple-choice format to ensure the consistency and certainty of the evaluation criteria. AncientBench contains 28,707 questions, all in multiple-choice format.

AncientBench Construction

Data Collection. The data sources for the AncientBench mainly include oracle bone inscription, Bronze Inscriptions, Bamboo Book of Chu, ShuoWen(Xu 1981), Chinese Dictionary Compendium(Luo 1986), and a series of pre-Qin books.

We have digitized excavated documents such as oracle bone inscriptions and Bamboo Book of Chu, as well as transmitted documents such as ShuoWen and Chinese Dictionary Compendium based on the constructed character coding list. Then we matched and integrated each character with its radicals, pronunciation, and character meaning information.

Digitization of Ancient Characters. One of the main difficulties in constructing datasets of ancient characters is that many ancient characters are not encoded in computers, and many ancient characters that can be represented in computers may have different unicode encoding. In order to standardize the evaluation of ancient characters for comprehension, digitization and encoding of ancient characters need to be standardized. In this paper, we use a three-phase approach to achieve digitization and encoding of ancient characters, i.e., ancient character image processing, unified font encoding, and new encoding of missing characters.

The first is ancient character image processing. We extract the feature information and spatial relationship information of each radical of the ancient character image by computer vision methods, and link the ancient characters with the radical information to generate the ancient character knowledge graph. Then we reconstruct a vector image for each ancient character.

The second is unified font encoding. We extract glyphs and encoding information from the font library based by character processing technology, and perform deduplication operations on characters to ensure that each character corresponds to a unique encoding information, and finally obtain an integrated character encoding table.

Finally the new encoding of missing characters. The glyph information obtained from the image processing of ancient characters is compared with the character encoding table, if it is an existing character, the unicode encoding in the character encoding table continues to be used, and if it is an unrecorded character, it is added to the vocabulary as a new character generating a new unicode encoding.

After the three-stage approach to digitization of ancient characters, we can obtain a complete and unified table for character encoding.

Subject Construction. Based on the four dimensions we defined for capability evaluation, we analyzed the existing available data resources, as well as referring to the classic tasks in the field of natural language reasoning, and finally identified ten evaluation tasks, i.e., Radical, Radical Meaning, Pronunciation, Phonetic Radical, Homophone, ExcDoc Word, TraDoc Word, Cloze, Phonetic Loan Character, and Translation. In order to ensure the accuracy and reasonableness of the topic construction, we followed the following principles in reconstructing the subject and options. Some sample questions are shown in Figure 2.

The first is uniform evaluation criteria. In order to standardize the input and output of the model, as well as ensuring uniformity in the criteria for evaluating the model, we used multiple-choice questions for all of them.

The second is difficulty level differentiation. We differentiated the difficulty levels of the AncientBench questions from three perspectives: task type, data source, and question option design. In terms of task type, ten evaluation tasks of AncientBench cover ancient characters from the understanding of various features of a single character to the semantic understanding in the context, so the Radical is defined as a relatively easy task, while the Cloze will be defined as a more difficult task. In terms of data sources, the AncientBench's data covers multiple eras such as oracle bones, Bronze Inscriptions, and Warring States scripts, etc. Scripts from the ancient period will have greater differences in the way of using words and in the glyph and pronunciation of words compared to those of more recent eras, so even though the ExcDoc Word belongs to the meaning comprehension ability, and the Translation belongs to the contextual comprehension ability, ExcDoc Word will be more difficult than Translation, because the words in ExcDoc Word is older than that in Translation. In the same task, different subjects also have difficulty differentiation, for example, the Radical includes Oracle Bone, Bronze Inscriptions, and Warring States scripts. Oracle Bone is more difficult to identify compared to Warring States script because the glyphs of Oracle Bone are quite different from modern Chinese characters. As for the design of the question options, in order to avoid the question options being too intuitive, we did not generate all the question options by random method, but carefully designed the difficulty differentiation. For some of the options, we replaced a radical of the correct answer with a similar radical to ensure the reasonableness of the questions, as shown in the red box in Figure 2, when we constructed option B, we chose the part of the answer that has a similar meaning to the correct answer as the confusing option. We further refine the distinction between models or human ancient characters by differentiating difficulty between tasks, and within tasks.

Finally the authority. After the AncientBench was constructed, we invited researchers in the field of archaeology to

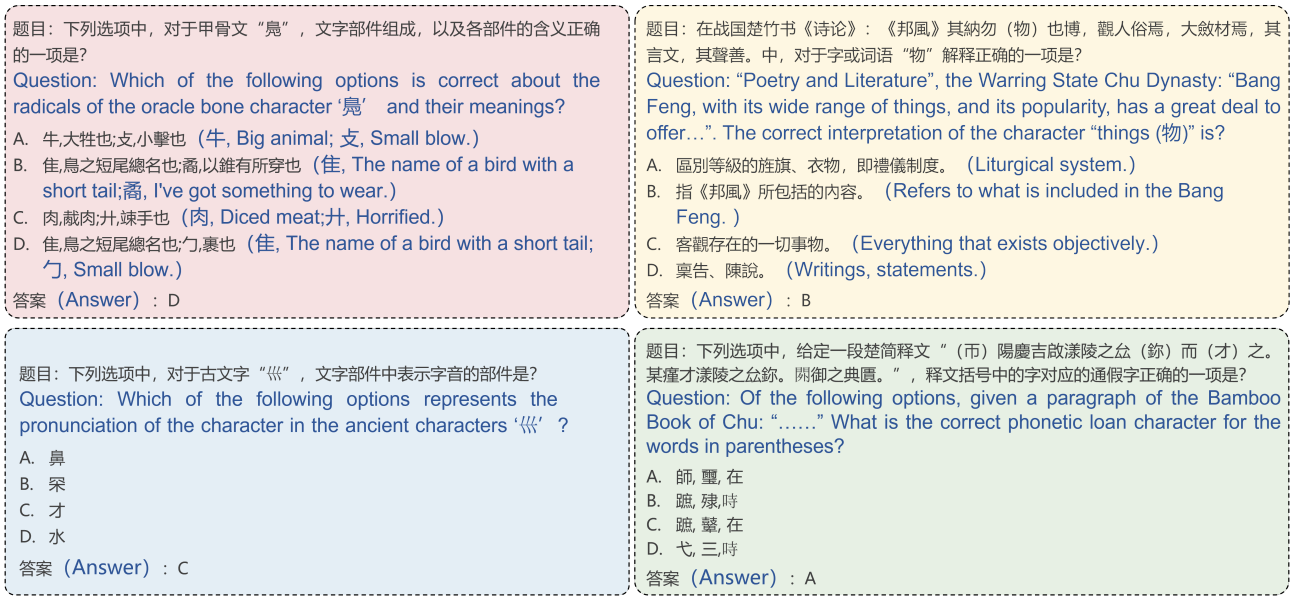


Figure 2: Some samples from the AncientBench. The red box shows a Radical Meaning task sample for glyph comprehension, the blue box shows a Phonetic Radical task sample for pronunciation comprehension, the yellow box shows an ExcDoc Word task sample for meaning comprehension, and the green box shows a Phonetic Loan Character task sample for contextual comprehension.

evaluate the AncientBench from the aspects of data source selection, topic design, and option design to ensure the reasonableness of the AncientBench.

Experiments

Setup

We evaluated both the LLMs and human comprehension of ancient characters based on AncientBench, and for the LLMs, we compared the accuracy of the models in both zero-shot and few-shot, respectively.

Referring to the experimental setup of CMMLU(Li et al. 2024), AC-EVAL(Wei et al. 2024), we used the following statement in constructing the model input: “For the following multiple-choice questions, please directly give the option of the correct answer”. For zero-shot, we will directly give the output prompt: “the answer is: ”. For few-shot, we will give five samples containing the correct answer.

Regarding the output of the model, we refer to the answer processing method of MMLU(Hendrycks et al. 2021a), the logits of the next of the next predicted token is obtained after inputting the prompt into the model, then we compare the probability of the four tokens, ‘A’, ‘B’, ‘C’ and ‘D’, and finally we choose the one with the highest probability as the model choice.

For the matrices, we calculated multiple-choice accuracy for the ten tasks, averaged the accuracy for the tasks included in the four competencies as the score for each competency, and finally calculated the average of all competency scores as the average score.

Human Performance. To facilitate comparison with the model, we evaluated the comprehension of ancient characters in humans. In conducting the evaluation, we randomly selected five multiple-choice questions for each of the 10 tasks, constituting a 50-question multiple-choice questionnaire, and then invited the participants to take part in the examination. We counted the participants’ accuracy and time to answer the questions, and took the average as the average of the accuracy of each task. The invited participants were graduate students at the intersection of the fields of archaeology and artificial intelligence.

Models. In our evaluation, we have selected 9 models that perform well in the Chinese field. Including (1) Llama-3-8B-Instruct(Grattafiori et al. 2024), (2) GLM-4-9b-Chat(Zeng et al. 2024a), (3) Qwen-7B/14B-Chat(Bai et al. 2023), (4) Baichuan2-7B/13B-Chat(Yang et al. 2025), (5) Yi-1.5-9B-Chat(Young et al. 2025), (6) Xunzi-Qwen-Chat(Zhao et al. 2024), (7) Tonggu-7b-Chat(Cao et al. 2024).

Ancient Model

In order to evaluate the impact of ancient Chinese knowledge on the ancient Chinese comprehension ability of LLMs, and to establish a strong baseline for benchmark, we fine-tuned LLMs based on ancient Chinese knowledge.

Based on our preliminary experiments, we used Yi1.5-9B-Chat(Young et al. 2025), which has the strongest ancient Chinese capabilities. We constructed a fine-tuning dataset. Specifically, we constructed fine-tuning question-answer pairs, each consisting of three parts: instruction, input, and output. The instruction describes what task to perform, the input describes the specific task details, and the output defines the model output. Based on this dataset, we performed full fine-tuning on Yi1.5-9B-Chat. During train-

Model	Glyph	Pronunciation	Meaning	Contextual	Average
human performance	76.66	50.00	38.33	55.55	55.13
Qwen-14B-Chat(Bai et al. 2023)	**53.01**	31.70	63.86	55.43	**51.00**
GLM4-9b-chat(Zeng et al. 2024a)	44.96	34.48	**69.71**	48.05	49.30
Baichuan2-13B-Chat(Yang et al. 2025)	45.08	**35.16**	60.81	56.11	49.29
Yi1.5-9B-Chat(Young et al. 2025)	42.70	30.12	65.53	56.91	48.81
Baichuan2-7B-Chat(Yang et al. 2025)	49.32	29.68	58.54	54.64	48.04
Llama3-8B-Instruct(Grattafiori et al. 2024)	41.54	28.15	51.80	**62.41**	45.97
Xunzi-Qwen-Chat(Zhao et al. 2024)	43.03	27.40	55.23	49.72	43.84
Qwen-7B-Chat(Bai et al. 2023)	45.80	25.08	56.33	48.14	43.83
Tonggu-7b-chat(Cao et al. 2024)	33.72	29.37	42.49	38.65	36.05
Yi1.5-9B-Ancient(ours)	50.77	32.30	67.12	50.93	50.28

Table 2: Accuracy of each model in zero-shot. We report the average of each tasks in comprehension capacity. "Average" is the average of the comprehension competencies. Bold indicates the best model performance.

ing, we set the batch size to 2 and the learning rate to $1e-5$. All experiments were executed on the machine running the Ubuntu OS with ascend-d910b npu. We named our model Yi1.5-9B-Ancient.

Zero-shot Results

We evaluated the comprehension of ancient characters by LLMs in zero-shot, and compared with human performance. We analyzed the results in terms of the number of parameters of the model, the type of model, and the different tasks.

The ancient characters comprehension of the LLMs in zero-shot is shown in Table 2. The average competence of ancient characters comprehension of LLMs is lower than the human performance, and even the best performing model, i.e., Qwen-14B-Chat, is 4.13% lower than the average human level. For glyph comprehension and pronunciation comprehension, there is a large difference between the LLMs and the human performance, and the human performance is generally more than 30% higher than the LLMs, which may be because humans are multimodal in recognizing ancient characters' glyphs, whereas the LLMs have only a single modality. For tasks such as context comprehension, which requires only a single modality, there is not much difference between the LLMs and the human performance. For meaning comprehension, the LLMs generally scored higher than the human performance, probably because the word sense comprehension task requires more memory.

From the perspective of model type, the average ancient characters comprehension competence of the ancient Chinese model and the generic LLMs did not differ much. In terms of the average score, Qwen-14B-Chat, Baichuan2-13B-Chat, and GLM4-9b-chat have the highest scores of 51.00%, 49.30%, and 49.29% respectively, which shows that although most of these models are trained on a modern Chinese corpus, they possess strong generalization ability. Llama3-8B-Instruct's is trained mainly on the English corpus, and thus scores slightly lower than the Chinese model with the same number of parameters, which is in line with our conjecture. With the same number of parameters, the glyph comprehension and pronunciation comprehension of the ancient Chinese model is slightly higher than that of the generic LLM, which may be because of the larger number of ancient characters included in the training corpus of the ancient Chinese model, which enables the model to build up information between the ancient characters and their glyph. The above observations reflect the linguistic differences between English and Chinese as well as between modern and ancient Chinese in the training corpora, further emphasizing the importance of our benchmarks.

From the perspective of the number of model parameters, the highest average score is Qwen-14B-Chat with 51.00%, which shows that increasing the number of model parameters is beneficial to the comprehension of ancient characters, the average score of Baichuan2-13B-Chat is 1.25% higher than that of Baichuan2-7B-Chat, and Qwen-14B-Chat's average score improved by 7.17% over Qwen-7B-Chat. For meaning comprehension and contextual comprehension, the effect of the number of parameters is more obvious, with Baichuan2-13B-Chat's meaning comprehension being 2.27% higher than Baichuan2-7B-Chat's, and Qwen-14B-Chat's contextual comprehension being 7.29% higher than Qwen-7B-Chat's, which may be because meaning comprehension and contextual comprehension require more logical reasoning skills compared to Glyph and Pronunciation comprehension.

From the perspective of AncientBench's different tasks, the LLMs scored high in the evaluation of meaning comprehension and contextual comprehension competence, with GLM4-9b-chat reaching 69.71% in meaning comprehension, which is 31.38% higher than the human performance; however, they performed poorly in the evaluation of glyph comprehension and pronunciation comprehension ability, with all the models scoring around 30 in the pronunciation comprehension evaluations, close to random selection. This may be because the questions give the context, and the models are able to use this information to reason the meanings of the ancient characters and then provide answers; whereas glyph comprehension and pronunciation comprehension directly require the models to recognize the glyph and pronunciation, which relies more on the amount of knowledge inherent in the models, i.e., the knowledge learned during the pre-training process. The above conclusion proves that the

Model	Glyph	Pronunciation	Meaning	Contextual	Average
human performance	76.66	50.00	38.33	55.55	55.13
Qwen-14B-Chat(Bai et al. 2023)	47.31 (-5.7)	29.75 (-1.95)	65.08 (+1.22)	58.10 (+2.67)	50.05 (-0.95)
GLM4-9b-chat(Zeng et al. 2024a)	46.33 (+1.37)	34.92 (+0.44)	**69.84** (+0.13)	51.05 (+3.00)	50.53 (+1.23)
Baichuan2-13B-Chat(Yang et al. 2025)	43.84 (-1.24)	**35.34** (+0.18)	57.13 (-3.68)	55.37 (-0.74)	47.92 (-1.37)
Yi1.5-9B-Chat(Young et al. 2025)	46.73 (+4.03)	32.04 (+1.92)	66.21 (+0.68)	58.26 (+1.35)	50.80 (+1.99)
Baichuan2-7B-Chat(Yang et al. 2025)	42.74 (-6.58)	32.10 (+2.42)	56.78 (-1.76)	50.54 (-4.1)	45.54 (-2.5)
Llama3-8B-Instruct(Grattafiori et al. 2024)	41.24 (-0.3)	27.10 (-1.14)	57.17 (+5.37)	**64.57** (+2.16)	47.51 (+1.54)
Xunzi-Qwen-Chat(Zhao et al. 2024)	44.63 (+1.6)	29.20 (+1.8)	52.67 (-2.56)	49.03 (-0.69)	43.88 (+0.04)
Qwen-7B-Chat(Bai et al. 2023)	42.53 (-3.72)	25.77 (+0.69)	55.47 (-0.86)	49.37 (+1.23)	43.28 (-0.55)
Tonggu-7b-chat(Cao et al. 2024)	34.54 (+0.82)	30.41 (+1.04)	45.34 (+2.85)	39.09 (+0.44)	37.34 (+1.29)
Yi1.5-9B-Ancient(ours)	**49.67** (-1.1)	32.98 (+0.68)	65.23 (-1.89)	55.44 (+4.51)	**50.83** (+0.55)

Table 3: Accuracy of each model in few-shot. We report the average of each tasks in comprehension capacity. "Average" is the average of the comprehension competencies. Bold indicates the best model performance.

reasoning ability and the amount of knowledge of LLMs can indeed assist us in the research in the field of ancient characters, but the current LLMs have less knowledge related to excavated documents, which is caused by the low content of excavated documents in the pre-training corpus of the models, which shows that the LLMs have a great potential in the research in the field of ancient Chinese, and also side by side proves the importance of our proposed benchmarks.

It is worth noting that Yi1.5-9B-Ancient, after being fine-tuned with our ancient Chinese knowledge, demonstrated excellent performance. The average score reached 50.28%, achieving the best performance among LLMs with the same parameter scale, and an average score increase of 1.47% compared to the Yi1.5-9B-Chat. In terms of glyph comprehension, Yi1.5-9B-Ancient achieved 50.77%, an improvement of 8.07% over Yi1.5-9B-Chat. In terms of pronunciation comprehension and meaning comprehension, Yi1.5-9B-Ancient achieved 32.30% and 67.12%, respectively. The above results prove that Yi1.5-9B-Chat contains relatively little information on ancient characters, and it is likely that LLMs such as Yi1.5-9B-Chat have limited embedding of ancient characters, while the method of fine-tuning can supplement this part of the knowledge of large language models. Yi1.5-9B-Ancient performed poorly on the contextual comprehension task, even 5.98% lower than Yi1.5-9B-Chat. This maybe because the contextual comprehension contains some topics related to traditional documents, and the ancient knowledge introduced by fine-tuning caused some damage to the original embedding of Yi1.5-9B-Chat, thereby affecting the model's performance on the contextual comprehension. In summary, existing LLMs lack knowledge related to ancient Chinese, resulting in weak ancient character comprehension abilities. Fine-tuning as a baseline is an intuitive and effective method that can supplement LLMs with ancient Chinese knowledge, but it may affect the original embedding. Therefore, it is necessary to further propose better methods that can represent ancient characters without affecting the original embedding.

Few-shot Results

We analyzed the few-shot results in terms of the number of parameters of the model, the type of model, and the different tasks, and compared the results with the zero-shot results.

The ancient characters comprehension of the LLMs in few-shot is shown in Table 3. Compared to zero-shot, the performance of most of the models is improved, but still lower than the human performance. Yi1.5-9B-Chat achieves a average score improvement of 1.99% to 50.80%. tonggu-7b-chat, glm-4-9b-chat, and yi-1.5-9B-Chat in few-shot have all the four competence scores all improved. However, both Baichuan2 and Qwen have reduced performance in few-shot, which is different from our expectation.

From the point of view of each task, the model's pronunciation comprehension and contextual comprehension improved in few-shot, and the model's pronunciation comprehension score in zero-shot was close to 25%, i.e., randomly selected, which may be because the model was unfamiliar with the content of the questions and the options, and the samples given in the model inputs in the few-shot experiments helped the model's comprehension, which led to the model's improved performance.

The Yi1.5-9B-Ancient model achieved 50.83% in few-shot, which is the best performance among all LLMs, an improvement of 0.03% over the Yi1.5-9B-Chat, and a 2.94% improvement in the glyph comprehension task score.

Conclusion

In this paper, we present AncientBench, a benchmark centered on excavated documents containing four competencies and ten tasks that comprehensively evaluates the comprehension of ancient characters for large language models. We propose a three-stage approach to digitize ancient characters and encode them into computers. In addition, we propose an ancient model as a baseline and evaluate human performance and large language models based on AncientBench. Our research introduces excavated documents into the field of natural language processing for the first time.

In future work, we will further expand the sources of data and extend it to multiple modalities. In addition, we will explore more effective model evaluation methods and metrics. Our research will provide a good foundation for large language models in the field of ancient Chinese.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.62476111), the Department of Science and Technology of Jilin Province, China (20230201086GX), the "Paleography and Chinese Civilization Inheritance and Development Program" Collaborative Innovation Platform (No.G3829), the National Social Science Foundation of China (No. 23VRC033), and the interdisciplinary cultivation project for young teachers and students at Jilin University, China (No. 2024-JCXK-04).

References

Austin, J.; Odena, A.; Nye, M. I.; Bosma, M.; Michalewski, H.; Dohan, D.; Jiang, E.; Cai, C. J.; Terry, M.; Le, Q. V.; and Sutton, C. 2021. Program Synthesis with Large Language Models. *CoRR*, abs/2108.07732.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; and Ma, J. 2023. Qwen Technical Report. arXiv:2309.16609.

Boltz, W. G. 1986. Early chinese writing. *World Archaeology*, 17(3): 420–436.

Bommasani, R.; Liang, P.; and Lee, T. 2023. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences*, 1525(1).

Cao, J.; Peng, D.; Zhang, P.; Shi, Y.; Liu, Y.; Ding, K.; and Jin, L. 2024. TongGu: Mastering Classical Chinese Understanding with Knowledge-Grounded Large Language Models. arXiv:2407.03937.

Chi, Y.; Giunchiglia, F.; Shi, D.; Diao, X.; Li, C.; and Xu, H. 2022. ZiNet: Linking Chinese Characters Spanning Three Thousand Years. In *Findings of the Association for Computational Linguistics: ACL 2022*, 3061–3070.

Conneau, A.; and Kiela, D. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In Calzolari, N.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Hasida, K.; and Isahara, H., eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA).

Diao, X.; Shi, D.; Li, J.; Shi, L.; Yue, M.; Qi, R.; Li, C.; and Xu, H. 2023a. Toward Zero-shot Character Recognition: A Gold Standard Dataset with Radical-level Annotations. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6869–6877.

Diao, X.; Shi, D.; Tang, H.; Shen, Q.; Li, Y.; Wu, L.; and Xu, H. 2023b. RZCR: Zero-shot Character Recognition via Radical-based Reasoning. In *IJCAI*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; and Yang, A. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Guiyuan, W. 2023. *A Study of Excavated Documents in China*. London: Routledge.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021b. Measuring Mathematical Problem Solving With the MATH Dataset. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Ho, C. S.-H.; Ng, T.-T.; and Ng, W.-K. 2003. A "radical" approach to reading development in Chinese: The role of semantic radicals and phonetic radicals. *Journal of literacy research*, 35(3): 849–878.

Jane, Q. 2014. Ancient times table hidden in Chinese bamboo strips.

Li, H.; Zhang, Y.; Koto, F.; Yang, Y.; Zhao, H.; Gong, Y.; Duan, N.; and Baldwin, T. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, 11260–11285.

Li, W.; Qi, F.; Sun, M.; Yi, X.; and Zhang, J. 2021. CCPM: A Chinese Classical Poetry Matching Dataset. arXiv:2106.01979.

Liu, C.; Jin, R.; Ren, Y.; Yu, L.; Dong, T.; Peng, X.; Zhang, S.; Peng, J.; and Zhang, P. 2023. M3KE: A Massive Multi-Level Multi-Subject Knowledge Evaluation Benchmark for Chinese Large Language Models. arXiv:2305.10263.

Liu, M.; Xiang, J.; Xia, X.; and Hu, H. 2022. Contrastive Learning between Classical and Modern Chinese for Classical Chinese Machine Reading Comprehension. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(2).

Luo, Z., ed. 1986. *Hanyu Da Cidian (The Grand Chinese Dictionary)*. Shanghai: Hanyu Da Cidian Press. Published in 12 volumes from 1986 to 1994.

Nelson, J. R.; and Stage, S. A. 2007. Fostering the development of vocabulary knowledge and reading comprehension though contextually-based multiple meaning vocabulary instruction. *Education and treatment of children*, 30(1): 1–22.

Pan, X.; Wang, H.; Oka, T.; and Komachi, M. 2022. Zuo Zhuan Ancient Chinese Dataset for Word Sense Disambiguation. In Ippolito, D.; Li, L. H.; Pacheco, M. L.; Chen, D.; and Xue, N., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, 129–135. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9): 99–106.

Shao, Y.; Shao, T.; Wang, M.; Wang, P.; and Gao, J. 2021. A Sentiment and Style Controllable Approach for Chinese Poetry Generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, 4784–4788. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384469.

Shi, D.; Diao, X.; Shi, L.; Tang, H.; Chi, Y.; Li, C.; and Xu, H. 2022a. CharFormer: A Glyph Fusion based Attentive Framework for High-precision Character Image Denoising. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Shi, D.; Diao, X.; Tang, H.; Li, X.; Xing, H.; and Xu, H. 2022b. RCRN: Real-world Character Image Restoration Network via Skeleton Extraction. In *Proceedings of the 30th ACM International Conference on Multimedia*.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *TRANSACTIONS ON MACHINE LEARNING RESEARCH*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Linzen, T.; Chrupała, G.; and Alishahi, A., eds., *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics.

Wang, P.; and Ren, Z. 2022. The Uncertainty-based Retrieval Framework for Ancient Chinese CWS and POS. In Sprugnoli, R.; and Passarotti, M., eds., *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 164–168. Marseille, France: European Language Resources Association.

Wei, Y.; Xu, Y.; Wei, X.; Yangsimin, Y.; Zhu, Y.; Li, Y.; Liu, D.; and Wu, B. 2024. AC-EVAL: Evaluating Ancient Chinese Language Understanding in Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1600–1617. Miami, Florida, USA: Association for Computational Linguistics.

Xu, L.; Hu, H.; Zhang, X.; Li, L.; Cao, C.; Li, Y.; Xu, Y.; Sun, K.; Yu, D.; Yu, C.; Tian, Y.; Dong, Q.; Liu, W.; and Shi, B. 2020a. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4762–4772.

Xu, L.; Hu, H.; Zhang, X.; Li, L.; Cao, C.; Li, Y.; Xu, Y.; Sun, K.; Yu, D.; Yu, C.; Tian, Y.; Dong, Q.; Liu, W.; Shi, B.; Cui, Y.; Li, J.; Zeng, J.; Wang, R.; Xie, W.; and Lan, Z. 2020b. CLUE: A Chinese Language Understanding Evaluation Benchmark. 4762–4772.

Xu, L.; Li, A.; Zhu, L.; Xue, H.; Zhu, C.; Zhao, K.; He, H.; Zhang, X.; Kang, Q.; and Lan, Z. 2023. SuperCLUE: A Comprehensive Chinese Large Language Model Benchmark. arXiv:2307.15020.

Xu, S., ed. 1981. *Shuowen Jiezi Zhu*. Shanghai: Shanghai Ancient Books Publishing House.

Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; Wang, D.; Yan, D.; Yang, F.; Deng, F.; Wang, F.; Liu, F.; Ai, G.; Dong, G.; Zhao, H.; Xu, H.; and Sun, H. 2025. Baichuan 2: Open Large-scale Language Models. arXiv:2309.10305.

Yang, H. 2000. A Brief Discussion on Phonetic Loan Characters. *Journal of Wuhan University (Humanities and Social Sciences Edition)*, 1.

Yao, Y.; Dong, Q.; Guan, J.; Cao, B.; Zhang, Z.; Xiao, C.; Wang, X.; Qi, F.; Bao, J.; Nie, J.; Zeng, Z.; Gu, Y.; Zhou, K.; and Huan, X. 2022. CUGE: A Chinese Language Understanding and Generation Evaluation Benchmark. arXiv:2112.13610.

Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; and Zhang, G. 2025. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652.

Yue, M.; Shi, D.; Diao, X.; Guo, S.; Li, C.; and Xu, H. 2025. Ancient character detection based on fine-grained density map. *npj Heritage Science*, 13(1): 280.

Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; et al. 2024a. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *CoRR*.

Zeng, H. 2023. Measuring Massive Multitask Chinese Understanding. *ArXiv*.

Zeng, H.; Xue, J.; Hao, M.; Sun, C.; Ning, B.; and Zhang, N. 2024b. Evaluating the Generation Capabilities of Large Chinese Language Models. arXiv:2308.04823.

Zhang, Y.; and Li, H. 2023. Can Large Language Model Comprehend Ancient Chinese? A Preliminary Test on ACLUE. In Anderson, A.; Gordin, S.; Li, B.; Liu, Y.; and Passarotti, M. C., eds., *Proceedings of the Ancient Language Processing Workshop*, 80–87. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.

Zhao, Z.; Shen, S.; Li, B.; and Ma, X. 2024. XunziALLM.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2024. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 2299–2314. Mexico City, Mexico: Association for Computational Linguistics.

Zhou, B.; Chen, Q.; Wang, T.; Zhong, X.; and Zhang, Y. 2023. WYWEB: A NLP Evaluation Benchmark For Classical Chinese. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 3294–3319. Toronto, Canada: Association for Computational Linguistics.

Zinin, S.; and Xu, Y. 2020. Corpus of Chinese Dynastic Histories: Gender Analysis over Two Millennia. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 785–793. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.